DOCUMENT RESUME

ED 443 838 TM 031 483

AUTHOR De Champlain, Andre F.; Gessaroli, Marc E.; Tang, K. Linda; De Champlain, Judy E.

TITLE Assessing the Dimensionality of Polytomous Item Responses with Small Sample Sizes and Short Test Lengths: A Comparison of Procedures.

PUB DATE 1998-04-14

NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Chi Square; Goodness of Fit; *Item Response Theory; *Sample Size; *Test Length

IDENTIFIERS *DIMTEST (Computer Program); *LISREL Computer Program; Polytomous Items

ABSTRACT

The empirical Type I error rates of Poly-DIMTEST (H. Li and W. Stout, 1995) and the LISREL8 chi square fit statistic (K. Joreskog and D. Sorbom, 1993) were compared with polytomous unidimensional data sets simulated to vary as a function of test length and sample size. The rejection rates for both statistics were also studied with two-dimensional data sets simulated to vary as a function of test length, sample size, and degree of correlation between latent traits. Severely inflated Type I error rates were obtained with the LISREL8 chi square statistic in all conditions, with the exception of the 10-item data sets simulated to contain 500 and 1,000 simulees. Poly-DIMTEST T-empirical Type I error probabilities were at or near nominal values for the three sample sizes examined. In addition, the performance of the latter statistic was unaffected by the manipulation of sample size. Rejection rates using the LISREL8 chi square fit statistic were high across all simulated two-dimensional conditions, although results were encouraging for 10-item data sets containing 500 or 1,000 simulees. It appeared that neither procedure worked well with samples of less than 500 examinees. Results do suggest that with samples of 500 examinees or more, the LISREL8 chi square statistic can be useful for assessment of dimensionality, but the Poly-DIMTEST T-statistic lacks the power needed to use with samples of fewer than 20 items. (Contains 3 tables and 38 references.) (SLD)

Assessing the Dimensionality of Polytomous Item Responses with Small Sample Sizes and Short Test Lengths: A Comparison of Procedures.

André F. De Champlain and Marc E. Gessaroli

National Board of Medical Examiners

K. Linda Tang and Judy E. De Champlain

Educational Testing Service

Paper presented at the meeting of the National Council on Measurement in Education

San Diego, California

Tuesday April 14, 1998

Assessing the Dimensionality of Polytomous Item Responses with Small Sample Sizes and Short Test Lengths: A Comparison of Procedures.

The assumption of unidimensionality is central to item response theory (IRT). Common IRT models assume that the probability of a correct response to a given item can be modeled as a function of a single person parameter (θ), usually interpreted as the proficiency underlying the item response matrix (Hambleton & Swaminathan, 1985). In practice this assumption is rarely met, given that an item response will often be dependent not only upon the hypothesized ability but also on several ancillary proficiencies.

A considerable body of research has been dedicated, over the past fifteen years, to developing indices and statistics to assess the underlying dimensionality of item response matrices (c.f. De Champlain & Gessaroli, *in press*, for a review). At present, indices and statistics based on Stout's concepts of *essential independence* and *essential dimensionality* (DIMTEST) have been shown to be useful for assessing the dimensionality of dichotomously-scored responses in several conditions (Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1991; 1994; Nandakumar & Stout, 1993; Stout, 1987; 1990). Similarly, the use of fit indices and statistics based on a nonlinear factor analysis (NLFA) of an item response matrix to assess the dimensionality of a given data set has also proven to be helpful with binary item responses (De Champlain, 1996; De Champlain & Tang, 1997; Gessaroli, 1994; Gessaroli and De Champlain, 1996; Hattie, 1984; 1985; McDonald & Mok, 1995). In addition, De Champlain and Gessaroli (1997) have shown that factor analytic models implemented in common software packages (e.g. PRELIS2/LISREL8 (Jöreskog & Sörbom, 1993a; 1993b)) are promising with respect to assessing the dimensionality of dichotomously-scored items.

The body of research focusing on the assessment of dimensionality for polytomously scored items is, however, more sparse (De Ayala, 1994; 1995). De Ayala (1994; 1995) has shown that the accuracy with which item and ability parameters can be estimated for both the graded-response and partial credit models is questionable in certain multidimensional conditions that were examined. In light of the case-specificity problem pervasive with performance assessments (Linn & Burton, 1994; Shavelson, Baxter, & Gao, 1993), identifying the nature of the composite would seem to be of the utmost importance.

In response to this issue, Li & Stout (1994; 1995) proposed a polytomous extension of their DIMTEST procedure (Poly-DIMTEST). The *T*-statistic, computed within the Poly-DIMTEST software, appeared to maintain low Type I error rates (close to the nominal value) with simulated unidimensional data sets. However, the power of the statistic in correctly rejecting unidimensionality with simulated two-dimensional data sets was low with small sample sizes (less than 1000) and short tests (less than 25 items). NLFA-based procedures and accompanying fit statistics have also been proposed for unidimensional and multidimensional polytomous item response models (Bartholomew, 1983; Christoffersson, & Gunsjö, 1996; Jöreskog, 1994; Muthén, 1984).

Although promising, little research has been undertaken to examine the behavior of these polytomous dimensionality assessment procedures in more realistic testing conditions. In particular, few investigations have focused on examining the Type I error rates and power of these statistics with small sample sizes and short test lengths. A large number of performance assessments are composed of very few items or tasks. Portfolios often contain no more than a dozen scoreable sections (Moss, Beck, Ebbs, Matson, Muchmore, Steele, & Taylor, 1992; Nystrand, Cohen, & Dowling, 1993). Similarly, performance assessments that are being considered for inclusion into the United States Medical Licensure Examination contain less than 20 scored tasks (Clauser, Subhiyah, Nungester, Ripkey, Clyman, & McKinley, 1995; De Champlain & Klass, 1997). A study examining the behavior of polytomous dimensionality assessment procedures with small sample sizes and short test lengths might therefore yield beneficial information for performance assessments administered within a variety of contexts for national as well as local examinations.

Purpose

The two primary objectives of this investigation were as follows:

- To estimate and compare the empirical Type I error rates of Poly-DIMTEST (Li & Stout, 1995) and the LISREL8 chi-square fit statistic (Jöreskog & Sörbom, 1993a; 1993b) with polytomous unidimensional data sets simulated to vary as a function of test length and sample size;
- To examine the rejection rates for both statistics with two-dimensional data sets simulated to vary as a function of test length, sample size, and degree of correlation between latent traits.

Methods

Unidimensional conditions

In the first part of this investigation, the empirical Type I error rates of both statistics were examined under various conditions. Unidimensional polytomous item response vectors were simulated using the generalized partial-credit IRT model (Muraki, 1992; 1997) which states that the probability of reaching a particular score category k (denoted P_{jk}) on item j is given by

$$P_{jk}(\theta)=\frac{\exp[\sum_{v=0}^{k} z_{jv}(\theta)]}{\sum_{c=0}^{K} \exp[\sum_{v=0}^{c} z_{jv}(\theta)]}, \qquad (1)$$

with

$$z_{jk}=a_j(\theta-b_{jk}),$$

and where

a_j = the item discrimination parameter for item j;

b_{jk} = the threshold (or step) parameter for item j and category k;

θ = the proficiency estimate.

Note that b_{jk} can be further decomposed additively into two parts:

b_j = the item location parameter, i.e., the overall difficulty of item j;

d_v = the relative difficulty of step v in comparison to other steps within item j.

In other words, the probability that a randomly selected simulee of ability level θ has of reaching score category k rather than $k-1$ can be estimated as a function of how well the item discriminates between test takers of varying ability as well as the difficulty level of the item and the difficulty associated with reaching a given step in comparison to other steps.

In addition, the unidimensional polytomous data sets were generated according to three sample sizes (250, 500 and 1000 simulees) as well as two test lengths (10 and 20 items).

In order to simulate realistic item responses, parameters for the data generation were selected from PARSCALE (Muraki & Bock, 1993) estimates obtained from a nationally administered standardized patient examination (SPX). These are presented in Table 1.

Insert Table 1 about here

The 20-item data sets were composed of two 10-item tests, i.e., the parameters used to simulated responses to items 1-10 were identical to those employed to generate responses to items 11-20. Also, note that the response variable contained five levels for items one through three whereas it included four levels for items four through 10. Proficiencies were randomly generated from a $N(0,1)$ distribution. Each cell of this 2 x 3 design (test length x sample size) was replicated 100 times for a total of 600 unidimensional data sets.

Two-dimensional conditions

In the second part of this investigation, two-dimensional polytomous item response vectors were simulated using a multidimensional extension of the generalized partial credit model (Muraki, 1992; 1997) given by

$$P_{jk}(\theta)=\frac{\exp[\sum_{v=0}^{k} z_{jv}(\theta)]}{\sum_{c=0}^{K}\exp[\sum_{v=0}^{c} z_{jv}(\theta)]}, \qquad (2)$$

with

$$z_{jk}(\theta)=\sum_{m=1}^{M} a_{jm}\theta_{m}+c_{jk},$$

and where

a_{jm} = a slope parameter of item j and the m-th (m=1,2,...M) latent trait dimension;

c_{jk} = an intercept parameter for item j and category k (k=1,2,...,K);

$\underline{\theta}$ = a proficiency vector.

These two-dimensional item response vectors were also simulated according to the same two sample sizes and two test lengths outlined in the previous section of the proposal as well as according to:

- *dimension dominance*: 50% of the items required knowledge of θ_1 only and the remaining items required knowledge of θ_2.

and

- *Inter-proficiency correlation*: 0.0, 0.3, and 0.6.

The parameters previously outlined with the unidimensional conditions were utilized in the two-dimensional simulations. As suggested by Muraki (1997), the intercept parameters corresponding to the threshold values outlined in Table 1 were obtained using the following formula,

$$c_{jk}=-a_{i}b_{jk} \qquad (3)$$

where c_{jk}, a_i and b_{jk} have been previously defined. Finally, proficiencies were randomly generated from a N(0,1) distribution. Each cell of this 2 x 3 x 3 design (test length by sample size by level of inter-proficiency correlation) was replicated 100 times for a total of 1800 two-dimensional data sets.

Analyses

Poly-DIMTEST was run and the powerful *T*-statistic (Nandakumar & Stout, 1993) was computed for the unidimensional and two-dimensional data sets using all default options. Given that the general DIMTEST procedure is well known, the reader is referred to other sources for computational details (Nandakumar & Stout, 1993; Nandakumar, Yu, Li, & Stout, 1995). Based on the results reported by Nandakumar, Yu, Li, and Stout (1995), which suggest that using simple Pearson correlations *in lieu* of the more theoretically appropriate polychoric correlations yields very similar Type I and rejection rates for the Poly-DIMTEST *T*-statistic, we decided, for simplicity's sake, to fit a linear factor analytic model to Pearson item correlations to select items for inclusion into the *AT1* subtest.

The asymptotic covariance matrix of the polychoric correlations was estimated for all data sets using PRELIS2 (Jöreskog & Sörbom, 1993a). PRELIS2/LISREL8 (Jöreskog & Sörbom, 1993a; 1993b) is a comprehensive structural equation modeling (SEM) package which allows the user to fit a confirmatory factor analytic model to a polytomous item response matrix via several estimation procedures. It is therefore possible to assess the fit of a one-factor (i.e., unidimensional) model to a data set prior to calibrating the item responses using an IRT model. Regardless of the procedure specified, the parameters of factor analytic models in LISREL are estimated so as to minimize the following fit function:

$$F=(s-\sigma)' \, W^{-1}(s-\sigma), \tag{4}$$

where

s = Sample item covariance matrix;

σ = Reproduced covariance matrix from the model parameters;

W = A weight matrix referred to as the *correct weight matrix*.

With polytomous responses, s usually corresponds to sample estimates of the threshold and polychoric correlations; σ contains the reproduced threshold and polychoric correlation values and W is a consistent estimator of the asymptotic covariance matrix of s.[4]

A chi-square goodness-of-fit statistic, provided in LISREL8 to aid in assessing model fit, is given by

$$\chi^2=(N-1)* \ Min(F), \tag{5}$$

where, *N* corresponds to the number of simulees in the sample and *Min (F)* is the minimum value of the fit function given in equation (4) for a specific model. This statistic is distributed asymptotically as a chi-square distribution with degrees of freedom equal to

$$.5(p)*(p+1)-t,$$

where *p* is equal to the number of items and *t* is the number of independent parameters estimated in the model. Chi-square statistic values were thus computed for all simulated unidimensional and two-dimensional data sets.

Regarding unidimensional data sets, a logit-linear analysis was undertaken to model the effects of test length and sample size as well as the interaction of both variables with respect to decision accuracy (i.e., the number of times the assumption of unidimensionality was accepted and rejected (Type 1 error)). For two-dimensional data sets, the effects of test length, sample size, degree of inter-proficiency correlation and the various interaction terms of the latter factors with respect to decision accuracy were also estimated via a logit-linear analysis. The logit-linear analyses were undertaken in a forward hierarchical fashion starting with the simplest main effect and progressing towards incrementally more complex models while adhering to the rule that higher-order effects are included in the model solely if the corresponding lower-order effects are also included. A model was deemed acceptable if its corresponding *p*-value exceeded 0.15. Effects with *z*-values greater than 2.00 were treated as statistically significant. For the sake of simplicity, significant associations in the logit-linear analyses are discussed only in light of the independent variable(s). For example, a significant decision accuracy by sample size by test length association in the logit-linear model would be referred to as the effect of sample size by test length. Finally, logit-linear analyses were undertaken separately for each statistic.

Results

Unidimensional conditions

The number of rejections of the assumption of unidimensionality, based on the Poly-DIMTEST *T*-statistic and LISREL8 χ^2 statistic, are shown for all simulated unidimensional conditions in Table 2. Due to software restrictions, it was not possible to compute *T*-statistics in conditions that contained only 10 items. *T*-statistics were thus estimated solely for the 20-item data sets.

Insert Table 2 about here

A nominal Type I error probability value of .05 was selected for all analyses. Empirical Type I error rates ranged from 0.03 (for *T*-statistic values based on data sets generated to contain 20 items and 500 simulees) to .99 (for LISREL8 chi-square values associated with data sets simulated to contain 20 items and 250 simulees). The results from the logit-linear analysis for the *T*-statistic indicate that a model solely containing the dependent variable "decision accuracy" was sufficient to adequately account for the empirical Type I error rates, $L^2(2) = 3.458, p=.177$. That is, empirical Type I error rates were not significantly affected by sample size. With respect to the LISREL8 χ^2 statistic, logit-linear analysis results indicate that a fully-saturated model, i.e., including all associations, is required to adequately account for the empirical Type I error rates, $L^2(0) = 0.000, p=1.000$. The proportion of incorrect rejections of unidimensionality for the 10-item data sets dropped from .22 (250-simulee data sets) to .10 (500-simulee data sets) and finally, .09 (1000-simulee data sets). On the other hand, empirical Type I error rates for the 20-item data sets dropped from .99 (250-simulee data sets) to .87 (500-simulee data sets) and finally, .49 (1000-simulee data sets).

Two-Dimensional conditions

Poly-DIMTEST *T*-statistic and LISREL8 χ^2 statistic rejection rates of the assumption of unidimensionality for all two-dimensional simulated conditions are shown in Table 3. Again, due to computational limitations, *T*-statistic calculations were restricted to 20-item data sets. Also, the same nominal Type I error probability was adopted (0.05).

Insert Table 3 about here

Rejection rates ranged from 3/100 (Poly-DIMTEST *T*-statistic for data sets simulated to contain 20 items, 250 simulees and an inter-proficiency correlation of .60) to 100/100 (LISREL8 χ^2 statistic for the data sets generated to contain 20 items and 250 simulees, irrespective of inter-proficiency correlation, as well as 20-item, 1000 simulee data sets simulated to have zero correlation between proficiencies).

The results from the Poly-DIMTEST *T*-statistic logit-linear analysis indicate that a fully-saturated model is needed to significantly account for the number of acceptances and rejections of the assumption of unidimensionality, $L^2(4) = 5.105$, $p=.277$. With respect to the "sample size by proficiency correlation" interaction, the proportions of rejections of the assumption of unidimensionality for 250-simulee data sets, were equal to .27, .22 and .03 for data sets where inter-proficiency correlation was respectively set at 0.00, 0.30 and 0.60. For 500-simulee data sets, these proportions were equal to .28, .27 and .16 for data sets simulated to respectively have inter-proficiency correlation values of 0.00, 0.30 and 0.60. Finally, with regard to 1000-simulee data sets, proportions of rejection rates of the assumption of unidimensionality dropped from .64 ($r\theta_1,\theta_2$=0.00) to .44 ($r\theta_1,\theta_2$=0.30), and finally .35 ($r\theta_1,\theta_2$=0.60). Logit-linear analysis results for the LISREL8 χ^2 statistic show that a model including all main effects in addition to the "test length by sample size" and "sample size by proficiency correlation" interactions was needed to adequately account for the observed proportions of acceptances and rejections of the assumption of unidimensionality, $L^2(8) = 3.916$, $p=.865$. Regarding the "test length by sample size" interaction, results show that the proportion of rejections of the assumption of unidimensionality increased for the 10-item data sets from 0.877 (250-simulee data sets) to 0.950 (500-simulee data sets), and finally 0.977 (1000-simulee data sets). However, these rates tended to be more stable for 20-item data sets as evidenced by proportion of rejection rates equal to 1.00, 0.977 and 0.990 for 250, 500, and 1000-simulee data sets, respectively. For the "sample size by proficiency correlation" interaction, the proportion of rejections of the assumption of unidimensionality for the 250-simulee data sets dropped from 0.950 ($r\theta_1,\theta_2$=0.00 and $r\theta_1,\theta_2$= 0.30) to 0.915 ($r\theta_1,\theta_2$= 0.60) whereas it varied from .975 ($r\theta_1,\theta_2$=0.00) to 0.955 ($r\theta_1,\theta_2$= 0.30) and finally, 0.960 ($r\theta_1,\theta_2$= 0.60) for data sets simulated to contain 500 simulees. Finally, for 1000-simulee data sets, proportions of rejections dropped from 1.000 to 0.985, and finally 0.965 when the degree of correlation between underlying proficiencies was respectively set at 0.00, 0.30 and 0.60.

Discussion

The re-emergence of performance assessments in education has enabled practitioners to measure types of behaviors not previously targeted by traditional means such as selected response items. The need for more "authentic" measures, however, does not preclude rigorous psychometric analyses. The assessment of dimensionality is central to both classical and modern test theories. At the most basic level, the validity of a score-based inference (what Messick, 1989 refers to as the *structural aspect* of construct validity) rests upon our knowledge of the underlying dimensional structure of an item response matrix. The need to better understand the structure of our data is therefore of the utmost importance. The dearth of research dedicated to the assessment of dimensionality with polytomous data is of particular concern given the popularity of alternative forms of assessment in education at the present time.

The findings obtained in the first part of this study focused on the performance of the Poly-DIMTEST *T*- and LISREL8 chi-square statistics with unidimensional data sets. Severely inflated Type I error rates were obtained with the LISREL8 χ^2 statistic in all conditions, with the exception of 10-item data sets simulated to contain 500 and 1000 simulees. Poly-DIMTEST *T*- empirical Type I error probabilities were at or near nominal values for the three sample sizes examined. In fact, empirical alpha values were within two standard errors of the nominal value (.05) for all conditions examined. In addition, the performance of the latter statistic was unaffected by the manipulation of sample size. These results are similar to those reported by Li and Stout (1994) and Nandakumar, Yu, Li, and Stout (1995) in their simulations. On the other hand, the interaction of test length and sample size impacted upon Type I error rates of the LISREL8 χ^2 statistic computations.

It should be pointed out, however, that the fit statistic provided in the LISREL8 packages is chi-square distributed as: :ntotically. It is quite likely that the empirical Type I error rates estimated with the LISREL8 chi-square fit statistic would adhere more closely to the nominal alpha level with sample sizes exceeding those that were simulated in the present investigation. In fact, the lower Type I error rate obtained with 10-item data sets containing 500 and 1000 examinees seems to support this point. However, these larger sample sizes might represent unrealistic testing situations ' many locally-based performance assessments.

Not surprisingly given the inflated Type I error probabilities reported with 20-item data sets, rejection rates obtained using the LISREL8 chi-square fit statistic were high across all simulated two-dimensional conditions. Results were encouraging, however, for 10-item data sets containing 500 or 1000 simulees. The number of rejections of the assumption of unidimensionality in these conditions was equal to or greater than 90/100, irrespective of the degree of inter-proficiency correlation. Nonetheless, the logit-linear analysis results indicate that several factors impact upon the rejection rates obtained with the LISREL fit statistic. Poly-DIMTEST *T*-statistic results show that the procedure lacks power in all conditions examined. Again, these findings are similar to those reported by Nandakumar, Yu, Li, and Stout (1995). Also, the proportions of rejection rates were affected by the interaction of sample size and degree of inter-proficiency correlation.

Several tentative recommendations can be made based on the results obtained in the present study. First, it appears as though neither procedure works particularly well with samples containing less than 500 examinees. In those particular conditions, the onus should probably be placed on a sound test development process to ensure that the examination is targeting the intended constructs. Sireci and Geisinger (1995) have provided interesting applications of multidimensional scaling to ensure content domain representation. These types of analyses might also prove to be beneficial for assessing the structure of performance assessments administered to small samples.

With 10-item data sets containing 500 or more examinees, results suggest that the LISREL8 χ^2 statistic can be quite useful for the assessment of dimensionality. The Poly-DIMTEST *T*-statistic simply lacks the power needed in order to recommend its use with data sets that contain less than 20 items. However, based on prior findings, it is still advised to use the latter statistic with assessments that contain more than 20 items and 1000 examinees.

Having said this, it is important to state that these findings should be interpreted in light of several caveats. First, the reported findings are highly dependent upon the conditions that were simulated and generalizations to other configurations should be undertaken cautiously, if at all. For example, the item parameters selected for the simulations obviously might not reflect all performance assessments. However, there is little reason to believe that these parameter estimates would differ from other clinical skills assessments. Second, it is important to re-emphasize that the purpose of this study was to examine the behavior of both statistics in several conditions that would hopefully allow us to gather practical information regarding both procedures. Obviously, additional simulations should be undertaken before making any definitive statements about the Type I and Type II error rates of both statistics. Finally, the results obtained with the LISREL8 chi-square and Poly-DIMTEST *T*-statistics were not unexpected. Inflated Type I error rates were reported in a study by De Champlain and Gessaroli (1997) that examined the behavior of the LISREL goodness-of-fit statistic with dichotomously-scored responses. Also, past research has clearly shown that the *T*-statistic does not function well with most of the conditions examined in our study. Nonetheless, given the popularity and usefulness of the DIMTEST package more generally, we felt it important to compare the performance of the LISREL8 chi-square statistic to that of the *T*-statistic.

It is hoped that the results obtained in this study will provide valuable information regarding the behavior of two promising polytomous dimensionality assessment procedures in conditions that approximate those found with clinical skills performance assessments in medicine. It is also hoped that this investigation will help to foster future research in the area of dimensionality assessment in general. Finally, and more importantly, more attention needs to be geared towards the development and application of psychological and statistical models that aptly capture the complex multidimensional structure of performance assessments.

References

Bartholomew, D.J. (1983). Latent variable models for ordered categorical data. *Journal of Econometrics*, *22*, 229-243.

Christoffersson, A., & Gunsjö, A. (1996). A short note on the estimation of the asymptotic covariance matrix for polychoric correlations. *Psychometrika*, *61*, 173-175.

Clauser, B.E., Subhiyah, R.G., Nungester, R.J., Ripkey, D.R., Clyman, S.G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement*, *32*, 397-415.

De Ayala, R.J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*, 155-170.

De Ayala, R.J. (1995). The influence of dimensionality on estimation in the partial credit model. *Educational and Psychological Measurement*, *55*, 407-422.

De Champlain, A. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, *33*, p. 181-201.

De Champlain, A., & Gessaroli, M.E. (*in press*). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education.*

De Champlain, A.F.,& Gessaroli, M.E. (1997, April). *An empirical comparison of two LISREL chi-square goodness-of-fit statistics and the implications for dimensionality assessment of item response data.* Paper presented at the meeting of the American Educational Research Association, Chicago, Il.

De Champlain, A.F., & Klass, D.J. (1997). Assessing the factor structure of nationally administered standardized patient examination. *Academic Medicine*, *72*, s88-s90.

De Champlain, A.F. & Tang, K.L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement*, *57*, 174-178.

Gessaroli, M.E. (1994). The assessment of dimensionality via local and essential independence: A comparison in theory and practice. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M.W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp.93-104). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Gessaroli, M.E., & De Champlain, A. (1996). Using an approximate chi-square statistic to test for the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33*, 157-179.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Hattie, J., Krakowski, K., Rogers, J.H., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*, 1-14.

Jöreskog, K.G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*, 381-389.

Jöreskog, K.G., & Sörbom, D. (1993a). *PRELIS2 user's reference guide*. Chicago: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (1993b). *LISREL8 user's reference guide*. Chicago: Scientific Software International.

Li, H.H., & Stout, W. (1994, April). *Assessment of unidimensionality for partial credit polytomous items: A modification of DIMTEST*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Li, H.H., & Stout, W.F. (1995, April). *A version of DIMTEST to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*, 5-8.

McDonald, R.P., & Mok, M. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education.

Moss, P.A., Beck, J.S.., Ebbs, C., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1992). Portfolios, accountability, and an interpretative approach to validity. *Educational Measurement: Issues and Practice, 11*, 12-21.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. (1997). *RESGEN: Item response generator*. Unpublished computer manual.

Muraki, E., & Bock, R.D. (1993). *PARSCALE: IRT based test scoring and item analysis*. Chicago, Il.: Scientific Software International.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*. 115-132.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R. (1994). Assessing the dimensionality of a set of item responses - Comparison of different approaches. *Journal of Educational Measurement, 31*, 17-35.

Nandakumar, R., & Stout, W.F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18*, 41-68.

Nandakumar, R.. Yu, F., Li, H.H., & Stout, W.F. (1995. September). *Assessing unidimensionality of polytomous data*. Unpublished manuscript. Neward, DE. University of Delaware.

Nystrand, M., Cohen, A.S., & Dowling, N.M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1*, 53-70.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability in performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Sireci. S.G., & Geisinger, K.F. (1995). Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement, 19*, 241-255.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Table 1

Polytomous Item Response Parameters (PARSCALE) Used in Simulations

		Item Step Parameters			
Item	*a*	b_1	b_2	b_3	b_4
1	0.134	3.636	4.803	-1.403	-7.037
2	0.177	4.683	-1.823	1.010	-3.869
3	0.154	9.706	-4.283	-1.270	-4.153
4	0.160	3.675	0.665	-4.340	——
5	0.341	1.746	0.804	-2.550	——
6	0.257	0.544	2.171	-2.715	——
7	0.275	3.289	-0.675	-2.614	——
8	0.242	1.708	1.350	-3.059	——
9	0.598	1.448	0.215	-1.663	——
10	0.404	-1.089	1.724	-0.635	——

Table 2

Number of Rejections of the Assumption of Unidimensionality per 100 Data Sets: Unidimensional Conditions

	10 items			20 items		
	N=250	N=500	N=1000	N=250	N=500	N=1000
Poly-DIMTEST *T*-statistic	—[1]	—	—	9	3	5
LISREL8 χ^2	22	10	9	99	87	51

[1] Due to Poly-DIMTEST restrictions, it was not possible to compute *T*-statistic values for 10-item data sets.

Table 3

Number of Rejections of the Assumption of Unidimensional ¬· per 100 Data Sets:

Two-Dimensional Conditions

Procedure	Test Length	Sample Size	Proficiency Correlation	Number of Rejections
Poly-DIMTEST	10 items	250 simulees	$r(\theta_1,\theta_2)=0.00$	—[2]
	10 items	250 simulees	$r(\theta_1,\theta_2)=0.30$	—
	10 items	250 simulees	$r(\theta_1,\theta_2)=0.60$	—
	10 items	500 simulees	$r(\theta_1,\theta_2)=0.00$	—
	10 items	500 simulees	$r(\theta_1,\theta_2)=0.30$	—
	10 items	500 simulees	$r(\theta_1,\theta_2)=0.60$	—
	10 items	1000 simulees	$r(\theta_1,\theta_2)=0.00$	—
	10 items	1000 simulees	$r(\theta_1,\theta_2)=0.30$	—
	10 items	1000 simulees	$r(\theta_1,\theta_2)=0.60$	—
	20 items	250 simulees	$r(\theta_1,\theta_2)=0.00$	27
	20 items	250 simulees	$r(\theta_1,\theta_2)=0.30$	22
	20 items	250 simulees	$r(\theta_1,\theta_2)=0.60$	3
	20 items	500 simulees	$r(\theta_1,\theta_2)=0.00$	28
	20 items	500 simulees	$r(\theta_1,\theta_2)=0.30$	27
	20 items	500 simulees	$r(\theta_1,\theta_2)=0.60$	16
	20 items	1000 simulees	$r(\theta_1,\theta_2)=0.00$	64
	20 items	1000 simulees	$r(\theta_1,\theta_2)=0.30$	44
	20 items	1000 simulees	$r(\theta_1,\theta_2)=0.60$	35

[2] Due to Poly-DIMTEST restrictions, it was not possible to compute *T*-statistic values for 10-item data sets.

Table 3 *(continued)*

Number of Rejections of the Assumption of Unidimensionality per 100 Data Sets:

Two-Dimensional Conditions

Procedure	Test Length	Sample Size	Proficiency Correlation	Number of Rejections
LISREL8 χ^2	10 items	250 simulees	$r(\theta_1,\theta_2)=0.00$	90
	10 items	250 simulees	$r(\theta_1,\theta_2)=0.30$	90
	10 items	250 simulees	$r(\theta_1,\theta_2)=0.60$	83
	10 items	500 simulees	$r(\theta_1,\theta_2)=0.00$	96
	10 items	500 simulees	$r(\theta_1,\theta_2)=0.30$	95
	10 items	500 simulees	$r(\theta_1,\theta_2)=0.60$	94
	10 items	1000 simulees	$r(\theta_1,\theta_2)=0.00$	100
	10 items	1000 simulees	$r(\theta_1,\theta_2)=0.30$	99
	10 items	1000 simulees	$r(\theta_1,\theta_2)=0.60$	94
	20 items	250 simulees	$r(\theta_1,\theta_2)=0.00$	100
	20 items	250 simulees	$r(\theta_1,\theta_2)=0.30$	100
	20 items	250 simulees	$r(\theta_1,\theta_2)=0.60$	100
	20 items	500 simulees	$r(\theta_1,\theta_2)=0.00$	99
	20 items	500 simulees	$r(\theta_1,\theta_2)=0.30$	96
	20 items	500 simulees	$r(\theta_1,\theta_2)=0.60$	98
	20 items	1000 simulees	$r(\theta_1,\theta_2)=0.00$	100
	20 items	1000 simulees	$r(\theta_1,\theta_2)=0.30$	98
	20 items	1000 simulees	$r(\theta_1,\theta_2)=0.60$	99